Positive-Unlabeled Learning for Addressing Hidden Positives in Survey-Based Health Screening Information Systems

Rafael F. Pinheiro[1], **Nataly L. Patti da Silva**[1]

[1]PPGSI-EACH
Universidade de São Paulo (USP)

{r.pinheiro,natalypatti}@usp.br

Abstract. *Survey-based health datasets embed label bias due to underdiagnosis and underreporting, undermining their use for predictive models in screening information systems. This paper explores Positive–Unlabeled (PU) Learning as a data-quality correction mechanism for self-reported health data. Using BRFSS-2015 and diabetes-related conditions, this paper shows how PU Learning can redistribute hidden positives within the unlabeled majority, improving detection of at-risk individuals—especially pre-diabetes—while shifting predictive signals toward healthcare access and response-quality factors. The results suggest that PU Learning can improve survey-based screening systems under incomplete labeling, challenging the no-label/no-condition assumption.*

1. Introduction

Survey-based datasets are widely used in health and social research due to their scalability and cost-effectiveness. They provide valuable input for prevention strategies, population monitoring, and decision-support systems. However, their reliability is limited by underdiagnosis and underreporting, which introduce bias into class labels and impair the performance of predictive models [Bekker and Davis 2020]. These issues are particularly critical in chronic disease research, in which hidden cases compromise both epidemiological estimates and the usefulness of machine learning for screening.

As an illustrative case, this study examines the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey data[1] to explore how underreporting affects predictive modeling in self-reported health datasets. When comparing reported pre-diabetes and diabetes cases in BRFSS to expected incidence rates in the U.S. population, a substantial degree of underreporting becomes apparent. The prevalence of reported diabetes was 12.97% versus an expected 14.7%, while pre-diabetes was reported in only 1.74% of respondents, compared to an expected 37.65% according to the National Diabetes Statistics Report 2021 (Centers for Disease Control and Prevention (CDC), 2021)[2]. Correspondingly, negative responses were overrepresented (84.29% in the dataset versus an estimated 47.65% non-diabetic population), indicating a systematic underdetection of at-risk individuals in self-reported survey data.

In this example, the observed discrepancy in pre-diabetes reporting may be a consequence of the survey's question design[3], as described in the BRFSS 2015 Codebook

[1]https://www.cdc.gov/brfss/annualdata/2015/files/LLCP2015XPT.zip

[2]https://www.cdc.gov/diabetes/php/data-research/index.html

[3]*"(Ever told) you have diabetes (If "Yes" and respondent is female, ask "Was this only when you were pregnant?". If Respondent says pre-diabetes or borderline diabetes, use response code 4.))*

Report[4], which requires respondents to proactively disclose a diagnosis of pre-diabetes. Moreover, this mechanism may amplify the existing underdiagnosis bias associated with these conditions, as reflected by CDC estimates showing that only 18.8% of pre-diabetic individuals are aware of their condition, compared to 77.95% among those with diabetes. Consequently, the BRFSS 2015 variable derived from this question (`DIABETE3`) may systematically underestimate the prevalence of pre-diabetes and, to a lesser extent, that of diabetes overall.

Traditional supervised learning methods struggle in this context. Class imbalance and the presence of hidden positives within the negative class reduce accuracy, recall, and generalization. Prior studies on diabetes have successfully applied supervised algorithms to survey data but typically assumed label completeness [Xie et al. 2019, Lakshmi et al. 2023, Alqahtani et al. 2024, Zhang et al. 2022, Zhang et al. 2020]. Nevertheless, this overlooks the structural biases of questionnaires and leads to the systematic underdetection of at-risk individuals.

To address these limitations, this study investigates whether Positive–Unlabeled (PU) Learning can mitigate hidden-positive bias [Elkan and Noto 2008, Mordelet and Vert 2014] in BRFSS-2015 diabetes screening data by redistributing unlabeled cases before supervised classification. Beyond its methodological contribution, the approach also opens space for a broader reflection within the Information Systems community. Conventional predictive models trained on self-reported health data implicitly assume that the absence of a label corresponds to the absence of a condition—an assumption rarely questioned, yet structurally embedded in large-scale survey-based information systems. This work challenges that premise by showing how such pipelines transform sociotechnical constraints (e.g., diagnostic access, questionnaire phrasing, response inequalities) into algorithmic ground truth, reinforcing invisibilities rather than merely reflecting them. By employing PU Learning not only as a technical mechanism but also as a critical lens, this study shows how survey-based information systems participate in the production of epidemiological blind spots. The findings demonstrate that the method not only improves predictive accuracy and aligns estimated disease prevalence more closely with epidemiological benchmarks, mitigating underreporting bias, but also invites the reader to reconsider how seemingly neutral modeling practices may inadvertently perpetuate detection asymmetries in vulnerable populations.

2. Related Work

While most predictive studies implicitly assume that negative labels correspond to true absence of risk, this assumption is often violated in large-scale surveys, where diagnosis, awareness, and disclosure are unevenly distributed across populations. In this context, PU Learning provides a principled framework for rethinking how labels produced by survey-based information systems should be interpreted and modeled.

PU Learning has attracted growing attention as a machine learning paradigm for situations in which only a fraction of positive cases are explicitly labeled, while the remaining positives are hidden among unlabeled or negative instances [Elkan and Noto 2008, Mordelet and Vert 2014]. This challenge arises across many application domains but is particularly critical in survey-based health research, in which

[4] `https://www.cdc.gov/brfss/annualdata/2015/pdf/codebook15llcp.pdf`

underdiagnosis and underreporting distort class distributions and reduce the reliability of supervised models.

To estimate the true probability of positivity $P(y = 1 \mid x)$ under subdiagnosis, a correction factor [Elkan and Noto 2008] was employed, expressed as:

$$\boldsymbol{P(y = 1 \mid x)} \approx \frac{P(s = 1 \mid x)}{c},$$

where:

- $\boldsymbol{P(s = 1 \mid x)}$ denotes the estimated probability that an instance is labeled as positive.
- $\boldsymbol{c = P(s = 1 \mid y = 1)}$ represents the fraction of true positives that were actually labeled.

The division by c compensates for the underestimation caused by incomplete labeling, and the adjusted values are constrained to the valid probability range $[0, 1]$.

Despite its suitability, PU Learning remains underexplored in large-scale survey applications. Most health analytics studies continue to rely on traditional supervised methods, even in the presence of pervasive mislabeling and class imbalance. For instance, prior work on diabetes and pre-diabetes prediction using BRFSS and similar surveys [Xie et al. 2019, Lakshmi et al. 2023, Alqahtani et al. 2024, Zhang et al. 2022, Zhang et al. 2020] employed standard classifiers and demonstrated the relevance of demographic, behavioral, and clinical features, while largely overlooking the impact of hidden positives.

PU Learning has nevertheless shown strong results in health-related and biomedical domains, including disease gene identification [Yang et al. 2012], drug–drug interaction prediction [Zheng et al. 2019], and other settings with incomplete annotations [Li et al. 2022, Bekker and Davis 2020]. These studies indicate that PU frameworks are well suited to learning under uncertain or noisy negative labels.

Underreporting is common in health surveys, particularly for socially sensitive conditions such as maternal smoking, which may be underestimated by up to 47% [Bekker and Davis 2020]. In this context, PU Learning provides a natural mechanism for correcting hidden positives and producing more realistic risk estimates.

Building on this gap, the present study demonstrates, using BRFSS-2015 diabetes data, how PU Learning can mitigate subdiagnosis bias and improve predictive performance in survey-based disease modeling.

3. Materials and Methods

3.1. Dataset

The experiments were conducted using the 2015 Behavioral Risk Factor Surveillance System BRFSS dataset, which includes 441,456 individual responses collected across all U.S. states and territories through structured telephone interviews. From the 330 original variables, a subset of 22 features was selected based on established associations with diabetes and pre-diabetes risk factors [Xie et al. 2019]. These features include demographic, behavioral, and self-reported clinical indicators.

Data preprocessing. Continuous variables were normalized to the [0,1] interval, with missing values replaced by a special value (-1). Ordinal variables were mapped to preserve their natural order, and categorical variables were one-hot encoded with additional binary flags for missing or refused responses. This encoding strategy tries to retain the semantics of nonresponse, here treated as a potential signal of underdiagnosis and access-related bias in survey-based information systems.

Class definitions. Target labels were derived from the BRFSS variable `DIABETE3`, which identifies respondents' diabetes self-report. Three classes were defined: no pre-diabetes or diabetes reported (`0`) ; pre-diabetes reported (`1`); diabetes reported (`2`). Importantly, the absence of a reported condition is interpreted as an uncertain label rather than confirmed absence of disease.

Dataset split and usage. The dataset was randomly partitioned to produce an 80/20 train-test division. This yielded 353,165 records for training and 88,291 for testing. The training subset was used both for PU Learning (estimation of the correction factors c and redistribution of hidden positives) and for training two classification models — one using the original labels and another using the PU-adjusted labels.

The test set was held out exclusively for evaluation and was not accessed at any stage of PU estimation or model fitting, thereby preventing any form of data leakage. This separation reflects a screening-oriented scenario, in which labeling correction and model training rely exclusively on historical data.

3.2. Positive-Unlabeled Learning

To address hidden positives among unlabeled instances, PU Learning was employed, as outlined in Algorithm 1. In this approach, pre-diabetes and diabetes were treated as positive labels in separate PU runs, whereas X_u denotes respondents with no reported condition.

Algorithm 1 PU Training and c_factor Estimation [Elkan and Noto 2008]

1: **function** TRAINPU(X_p, X_u)
2: Split X_p into (X_p^{train}, X_p^{val}) with 80/20
3: $X^{PU} \leftarrow X_p^{train} \cup X_u$
4: $y^{PU} \leftarrow [1 \text{ for } X_p^{train}] \cup [0 \text{ for } X_u]$
5: Train classifier clf_{pu} on (X^{PU}, y^{PU})
6: $\hat{p} \leftarrow clf_{pu}$.predict_proba$(X_p^{val})[:, 1]$
7: $c \leftarrow \text{mean}(\hat{p})$
8: **return** (c, clf_{pu})

Lines 1–3 isolate a held-out subset of labeled positives and combine the remaining positives with the unlabeled pool to form the PU training set; lines 4–6 train clf_{pu} to distinguish labeled from unlabeled cases and estimate c as the mean predicted score over the held-out positives.

In addition, pre-diabetes and diabetes reports were treated as positive labels separately and a 5-fold cross-validation was employed, where Algorithm 1 was executed independently on each fold, and the final output c was computed from the results obtained across all folds, resulting in the following means and standard deviations:

$$c_{\text{pre-diabetes}} = 0.5052 \pm 0.0080, \quad c_{\text{diabetes}} = 0.6674 \pm 0.0016.$$

These estimates represent the mean labeling propensity of pre-diabetes and diabetes among truly positive cases in the survey. Inference follows the standard PU correction, rescaling predicted labeling probabilities by the estimated c factor. This formulation assumes that labeled positives constitute a representative subset of true positives, while non-reported cases are treated as unlabeled rather than confirmed negatives.

Applying this correction to the unlabeled population produced a new inferred distribution of classes (Table 1), aligning the sample more closely with epidemiological expectations for the U.S. adult population.

Table 1. Class distributions before and after PU Learning

Class	Original (%)	After PU Learning (%)
Negative	85.1	35.1
Pre-diabetes	1.7	45.9
Diabetes	13.1	19.0

Crucially, this alignment is achieved without incorporating any external epidemiological or population data — no prevalence priors, post-stratification weights, or demographic adjustments were used. The estimates rely solely on questionnaire responses by modeling and applying the class-specific correction factors derived from the labeled positives; external statistics are referenced only for ex-post validation. These estimates should be interpreted as proxies of latent risk under incomplete labeling, rather than as direct measurements of population prevalence.

Table 2 reports the *row-normalized* confusion between the original labels (y) and the PU Learning labels (label_{pu}) on the test dataset. Each row shows the conditional distribution $P(\text{label}_{pu} \mid y)$. The relabeling reallocates a sizable share of $y{=}0$ (negative) into $\text{label}_{pu}{=}1$ (pre-diabetes), whereas $y{=}1$ and $y{=}2$ (pre-diabetes and diabetes, respectively) are preserved and remain concentrated on their respective classes. Similar redistribution patterns were observed in both training and test datasets.

Table 2. Original labels (y) vs. PU labels (label_{pu}) on the test dataset.

$y \setminus \text{label}_{pu}$	Negative	Pre-diabetes	Diabetes
Negative	49.00%	45.39%	5.61%
Pre-diabetes	0.00%	100.00%	0.00%
Diabetes	0.00%	0.00%	100.00%

3.3. Machine Learning Models

All classification models were implemented using the XGBoost algorithm, chosen for its well-established robustness for heterogeneous tabular data, and utilizing class weights inversely proportional to their frequency to handle class imbalance. This design isolates the effect of PU-based label redistribution, as the classifier is kept fixed; the goal was not model benchmarking, but a focused assessment of how PU Learning mitigates label incompleteness in screening-oriented information systems. Two model configurations were evaluated:

- **XGBoost:** baseline supervised model trained with original survey labels;
- **XGBoost_PU:** supervised model trained with redistributed labels from PU Learning.

All models shared identical hyperparameters to ensure comparability: `random_state = 1` and `eval_metric = "logloss"`. Training followed a 5-fold cross-validation.

3.4. Evaluation and Metrics

Model performance was primarily assessed using the Macro F1-score, which provides balanced sensitivity across all classes regardless of prevalence. Complementary metrics — accuracy, precision, recall, and area under the ROC curve (AUC-ROC) per class — were also computed to assess both overall and class-specific discrimination.

Training results are reported as mean ± standard deviation across five folds; final results come from models trained on the full training set and evaluated on the test set.

Particular emphasis was placed on recall for the positive classes (pre-diabetes and diabetes), given their relevance to early detection and health surveillance under incomplete labeling conditions.

3.5. Feature relevance and stability

Feature relevance evaluations were performed using the `BorutaPy` package. Boruta algorithm is a wrapper-based feature selection method that identifies statistically supported relevant features using Random Forests [Kursa and Rudnicki 2010]. In this study, Boruta was used as an interpretability tool rather than as a feature reduction mechanism.

Boruta was adopted to identify differences in relevant features between the classic and the PU-adjusted scenarios as an approach to reveal whether PU Learning can influence latent pattern recognition when applied as a pre-processing step to supervised machine learning classification tasks. Changes in feature relevance are interpreted as shifts in the signals captured by the model once hidden positives are made explicit.

3.6. Experimental Environment and Reproducibility

All experiments were executed in Python 3.8.7 using the following core libraries and versions: `NumPy 1.22.4`, `Pandas 2.0.1`, `Scikit-learn 1.1.1`, `XGBoost 1.6.1`, and `BorutaPy 0.4.3`. A fixed random seed (`random_state = 1`) was used across all experiments to ensure deterministic behavior.

4. Results

4.1. Comparative Classification Metrics

Initially, during cross-validation using the train dataset, the hidden positives limited the predictive accuracy of the baseline model. With PU Learning, however, performance improved substantially, supporting its potential usefulness for triage, as indicated in Tables 3.

Table 3. Preliminary performance on cross-validation

Model	Accuracy		Macro F1	
	Mean	Std	Mean	Std
XGBoost	0.65	0.003	0.42	0.002
XGBoost_PU	0.80	0.001	0.77	0.001

When retrained on the full training dataset and evaluating the test dataset, the baseline model's performance confirmed that conventional supervised learning alone struggles with biased survey labels, while the PU-adjusted model outperformed in terms of accuracy and macro F1, as shown in Table 4.

Table 4. Final performance on the test data

Model	Accuracy	Macro F1
XGBoost	0.65	0.42
XGBoost_PU	0.81	0.78

In fact, when precision, recall, and F1-score were evaluated on a per-class basis, all metrics showed substantial improvement under the PU learning model, except for precision in the negative class, which slightly decreased by six percentage points, as detailed in Table 5.

Table 5. Classification metrics by class on test data

Model	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
XGBoost	Negative	0.96	0.66	0.78	74,416
	Pre-diabetes	0.03	0.25	0.05	1,591
	Diabetes	0.33	0.64	0.43	11,403
XGBoost_PU	Negative	0.90	0.90	0.90	36,463
	Pre-diabetes	0.79	0.76	0.77	35,368
	Diabetes	0.64	0.68	0.65	15,579

To complement the within-label evaluation previously reported, Table 6 presents a cross-label comparison, in which both models are evaluated against both the original labels (y_{test}) and the PU-adjusted labels (y_{test}^{PU}). While the previous results preserve consistency between the labels used for training and testing, this crossed analysis highlights how each model behaves under alternative labeling assumptions, particularly with respect to hidden positives.

Table 6. Positive recall and false–positive rate (FPR) on test data

Model	Recall (vs. y_{test})	Recall (vs. y_{test}^{PU})	FPR (vs. y_{test})	FPR (vs. y_{test}^{PU})
XGBoost	0.83	0.68	0.34	0.03
XGBoost_PU	0.92	0.93	0.52	0.10

Compared to the baseline, the PU–adjusted model markedly increases sensitivity to at–risk cases. Against the original labels (y_{test}), recall improves by five percentage points, whereas against the PU–adjusted labels (y_{test}^{PU}) — a proxy for hidden positives — the gain is even larger, by 25%. This pattern is consistent with PU learning uncovering positives previously embedded in the unlabeled majority.

The trade–off is a higher false–positive rate: by 18% versus y_{test}, and by seven percentage points versus y_{test}^{PU}. From an information systems perspective, this trade-off reflects a deliberate shift toward sensitivity in screening contexts, where false negatives are typically more costly than false positives; however, it should be accompanied by threshold tuning and cost–sensitive evaluation to balance recall gains against the operational cost of additional false alarms.

4.2. AUC-ROC Performance

The AUC-ROC results reinforce the improvements obtained with PU Learning, providing a threshold-independent assessment of class separability. AUC-ROC is reported in Table 7 as a complementary indicator.

Table 7. AUC-ROC per class (without and with PU Learning)

Model	Negative	Pre-diabetes	Diabetes
XGBoost	0.82	0.67	0.83
XGBoost_PU	0.97	0.91	0.90

Table 7 presents the AUC-ROC of the baseline model using the original labels and the PU-adjusted model using the redistributed labels from PU Learning. The gains observed for all classes in the AUC-ROC metrics indicate that redistributing hidden positives provides a more reliable decision boundary, enhancing the robustness of the classifier for screening purposes.

Notably, the increase in AUC-ROC for the negative class by 15 percentage points is particularly informative. In a one-vs-rest ROC formulation, a higher negative-class AUC indicates that potential truly negative respondents are now much more cleanly separated from the rest, yielding higher specificity for a broad range of thresholds. From a screening perspective, this behavior can reduce unnecessary follow-ups among genuinely low-risk individuals while preserving sensitivity to at-risk cases. These gains reflect improved separability under incomplete labeling, not diagnostic performance at a fixed clinical threshold.

4.3. Confusion Matrices

Confusion matrices qualitatively illustrate how PU Learning reshapes misclassification patterns. Table 8 presents the results of the baseline model using the original labels,

while Table 9 shows the results of the PU-adjusted model using the redistributed labels from PU Learning.

Table 8. Confusion matrix for baseline model (original labels)

True label	Predicted label		
	Negative	**Pre-diabetes**	**Diabetes**
Negative	49061	11136	14219
Pre-diabetes	470	413	708
Diabetes	1788	2294	7321

Table 9. Confusion matrix for the PU-adjusted model (adjusted labels)

True label	Predicted label		
	Negative	**Pre-diabetes**	**Diabetes**
Negative	32793	3120	550
Pre-diabetes	2875	26987	5506
Diabetes	848	4113	10618

With PU Learning, class separation becomes substantially more balanced. Negatives are more cleanly isolated (90% correctly classified), while pre-diabetes detection improves markedly, with 76% correctly classified. Diabetes recognition also increases to 68%, with reduced leakage into the negative class.

Taken together, the confusion matrices confirm that the PU-adjusted model reduces cross-class confusion by reassigning potential hidden positives that were previously embedded in the negative class. This pattern supports the interpretation that many self-reported negatives correspond to unlabeled positives rather than true absence of risk, strengthening the reliability of survey-based screening systems.

4.4. Feature relevance and stability

Boruta feature selection [Kursa and Rudnicki 2010] identified 45 relevant features in the PU-adjusted scenario, compared to 20 under conventional supervised learning, indicating that PU Learning exposes additional latent patterns beyond classical risk factors. Full feature lists and descriptions are provided as supplementary material. From an information systems perspective, these shifts reflect how labeling mechanisms shape the signals available for decision support.

Without PU, relevant features are largely restricted to well-established biomedical and demographic predictors of diabetes risk, such as cardiovascular comorbidities, obesity, hypertension, and self-reported general health, consistent with prior literature.

With PU Learning, relevance expands to include indicators related to healthcare access, lifestyle, and systematic nonresponse (e.g., missing or refused answers). The prominence of these meta-features suggests that PU Learning enables the model to capture latent signals associated with underreporting, recall bias, and socioeconomic barriers.

These variables are interpreted as indicators of reporting dynamics rather than causal determinants.

Overall, the results indicate that PU Learning not only improves predictive performance, but also broadens interpretability by reducing over-reliance on a narrow set of classical risk factors and revealing sociotechnical dimensions of survey-based health data.

5. Conclusion

This study investigated whether PU Learning can address underdiagnosis and underreporting in survey-based health datasets. By redistributing unlabeled cases and approximating a more realistic class distribution, PU Learning acts as a corrective mechanism that enhances both the performance and interpretability of predictive systems derived from self-reported data. For Information Systems, the main contribution is to show that label incompleteness in survey-based systems should be treated as a structural property, not as incidental noise.

Future research may extend this framework by validating PU Learning across multiple survey waves, applying it to additional conditions affected by underreporting, and exploring its integration into broader decision-support architectures. PU Learning offers a promising pathway for transforming large-scale self-reported surveys into more accurate, equitable, and scalable predictive tools. At the same time, the results suggest that improving algorithms alone is not sufficient: it is also necessary to revisit the assumptions underlying what counts as evidence, absence, and uncertainty in Information Systems. If labels are treated as direct observations, rather than as sociotechnical products shaped by diagnosis access, questionnaire design, and reporting dynamics, supervised models will inevitably reproduce the absences and asymmetries embedded in them. In this sense, PU Learning is not only a methodological correction, but also an analytical device that makes visible what remains hidden under the negative label.

This leads to a broader provocation for the Information Systems community: how many other phenomena remain invisible because we accept, without scrutiny, the implicit ontology of labels in our systems? What opportunities for fairness, precision, and sociotechnical redesign emerge when the "not reported" is treated as an analytical object rather than as a negation? By reconciling methodological rigor with real-world data imperfections, this study advances both the technical and socio-organizational dimensions of Information Systems research and encourages further inquiry into how information systems produce and may help repair structural invisibilities.

Supplementary Material

The supplementary material — including preprocessed datasets, original labels, PU-adjusted labels, trained models, and Boruta feature lists — is openly available through an anonymous repository to preserve double-blind review.[5]

Acknowledgments

Language refinement of this manuscript was partially assisted by a generative AI model (OpenAI GPT-5). The methodological design, experiments, and interpretations were entirely conceived and validated by the authors.

[5] https://osf.io/bc3se/files/osfstorage?view_only=177811c4afd742ec91ca7c1293291519

References

Alqahtani, S. A. M., Alobaid, H. M., Alshammari, J., Alqarzae, S. A., Aloyouni, S. Y., Al-Eidan, A. A., Alhamad, S., Almiman, A., Alkhulaifi, F. M., and Alomar, S. (2024). Feature importance and model performance for prediabetes prediction: A comparative study. *Journal of King Saud University - Science*, 36:103583.

Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4):719–760.

Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 213–220, New York, NY, USA. Association for Computing Machinery.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13.

Lakshmi, H., Reddy, A. S., and Naidu, K. (2023). Analysis of diabetic prediction using machine learning algorithms on brfss dataset. In *Proceedings of the 7th International Conference on Trends in Electronics and Informatics (ICOEI 2023)*. IEEE.

Li, F., Dong, S., Leier, A., Han, M., Guo, X., Xu, J., Wang, X., Pan, S., Jia, C., Zhang, Y., Webb, G. I., Coin, L. J. M., Li, C., and Song, J. (2022). Positive-unlabeled learning in bioinformatics and computational biology: a brief review. *Briefings in Bioinformatics*, 23(1):bbab461.

Mordelet, F. and Vert, J.-P. (2014). A bagging svm to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209. Partially Supervised Learning for Pattern Recognition.

Xie, Z., Nikolayeva, O., Luo, J., and Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques. *Preventing Chronic Disease*, 16:190109. Original Research — Peer Reviewed.

Yang, P., Li, X., Mei, J., Kwoh, C., and Ng, S. (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647.

Zhang, L., Shang, X., Sreedharan, S., Yan, X., Liu, J., Keel, S., Wu, J., Peng, W., and He, M. (2020). Predicting the development of type 2 diabetes in a large australian cohort using machine-learning techniques: Longitudinal survey study. *JMIR Medical Informatics*, 8(7):e16850.

Zhang, P., Fonnesbeck, C., Schmidt, D. C., White, J., Kleinberg, S., and Mulvaney, S. A. (2022). Using momentary assessment and machine learning to identify barriers to self-management in type 1 diabetes: Observational study. *JMIR mHealth and uHealth*, 10(3):e21959.

Zheng, Y., Peng, H., Zhang, X., Zhao, Z., Gao, X., and Li, J. (2019). Ddi-pulearn: A positive-unlabeled learning method for large-scale prediction of drug-drug interactions. *BMC Bioinformatics*, 20(19):661.

Author Biographies

Rafael F. Pinheiro is an M.Sc. student in Information Systems at the University of São Paulo (USP). He holds a specialization in Data Science and Analytics and a degree in Computer Engineering, and has professional experience in data science, data engineering, analytics, and technology regulation. His work focuses on data-driven solutions, machine learning, and cloud-based data systems. His research interests include Information Systems, machine learning, data engineering, and data-intensive problem solving.

Nataly L. Patti da Silva holds both an M.Sc. and a B.Sc. in Information Systems from the University of São Paulo (USP) and is currently taking doctoral-level coursework in the same field at the same institution. She has professional experience in natural language processing, machine learning, and applied data science, and currently works as a Senior Data Scientist. Her research interests include Artificial Intelligence, computational methods for data and language analysis, and their applications in Information Systems.